

Building a WordNet for Dravidian Languages

S. Rajendran
Department of Linguistics
Tamil University, Thanjavur
raj_ushush@yahoo.com

G.Shivapratap, V.Dhanalakshmi, KP. Soman
CEN, Amrita Vishwa Vidyapeetham
Coimbatore, India
{g_shivapratap, v_dhanalakshmi, kp_soman}
@ettimadai.amrita.edu

Abstract

This paper attempts to emphasize the need for a standalone and independent Dravidian WordNet. Since the morphology and lexical concepts of Dravidian languages are closer to each other than to a language from a different family, it is proposed to base the Dravidian WordNet on a Dravidian Language. A significant amount of work has already been done in Tamil language to understand the ontological structure and vocabulary. Based on the findings of these studies, it is proposed to build a Tamil WordNet first and then extend it to complete the Dravidian WordNet. A prototype model for the Tamil WordNet is also proposed in this paper.

1 Introduction

The current work in Machine translation from English to Indian languages and between Indian languages demands strong lexical knowledge and resources. WordNet (Fellbaum, 1998) emerges as one of the natural and indispensable tool for this cause. The WordNet by its nature turns to be an ideal lexical accessing system as it links concepts with another concept by multifarious meaning relations. WordNet not only links one concept with another concept through semantic relations, but also captures the contextual meaning variations of a particular word i.e. the polysemy of a word.

Analogous to EuroWordNet, building an Indo WordNet needs individual WordNets for all the major Indian languages. Currently, WordNets are being developed at IIT-Bombay (Hindi and Marathi), Gujarati University, Ahmedabad (Gujarathi), IIT-Kharagpur (Bengali), and Tamil University, Thanjavur in collaboration with Amrita University and Kuppam University (Dravidian languages).

Among Indian languages, Dravidian languages such as Tamil, Telugu, Kannada and Malay-

lam share a number of lexicalized concepts in terms of morphology and semantics besides others as in typological and culture-specific features. The authors firmly believe that building a common WordNet for Dravidian languages will make it easier in developing an Indo-WordNet.

2 Need for Dravidian Wordnet

Dravidian WordNet is a natural chunk in the Indo-WordNet. It is only ideal that we should have Dravidian WordNet before we develop a larger Indo-WordNet, because the genealogical relationship among the Dravidian languages can be maximally exploited in a more natural way. It allows, for example, a search tool to infer other terms, from the terms provided by the user and coming up with the most optimal search for retrieving information.

Currently, the IndoWordNet is built using the Hindi wordnet as the source language. This approach is simpler and economical considering the interlinking of synsets of different languages. However, using source words from a different language family as a base to create synsets for Dravidian languages will restrict the syntactic envelope of the words in Dravidian languages. Hence, in order to capture the entire semantics, morphology and lexicalized concepts in Dravidian languages, we propose to build an independent WordNet for Dravidian languages.

3 Design of Dravidian WordNet

The design and implementation of the Dravidian WordNet will be based on EuroWordNet as explained by Pike Vossen (1998). The WordNet database will be built (as much as possible) from available existing resources and databases with semantic information developed in various projects. This will be not only more cost-effective given the limited time and budget of the project, but also will make it possible to combine information from independently created Word-

Nets. Two models (Vossen, 1999) will be involved in the built up.

- **Merge Model:** the selection will be done in a local resource and the synsets and their language-internal relations will be first developed separately, after which the equivalence relations to Tamil WordNet will be generated.
- **Expand Model:** the selection will be done in Tamil WordNet and the Tamil WordNet synsets will be translated (using bilingual dictionaries) into equivalent synsets in the other language. The WordNet relations will be later on adopted across languages.

The Merge Model will result in a WordNet that will be independent of Tamil WordNet, possibly maintaining the language-specific properties. The Expand model will result in a WordNet that is very close to Tamil WordNet but which is also biased by it.

3.1 Dravidian WordNet as extension of Tamil WordNet

The design of the Dravidian WordNet-database will be first of all based on the ontological structure of Tamil WordNet which in turn is based on a thesaurus for Tamil prepared by Rajendran (2001). Tamil WordNet relies on extensive preliminary investigations of the vocabulary of Tamil (Rajendran, 1976-2003) based on the componential analysis of meaning (Nida, 1975a & 1975b) and structural semantics (Lyons, 1977). Portions of this work have been compiled into a Tamil thesaurus (Rajendran, 2001). The Tamil thesaurus in electronic form represents the Ontological Structure of Tamil (OST) vocabulary giving scope to take care of any kind of semantic/lexical relations that hold between lexical items. The notion of a synset and the main semantic relations will be taken over in Dravidian WordNet. However, some specific changes will be made to the design of the database, which are mainly motivated by the following objectives:

- to create a multilingual database;
- to maintain language-specific relations in the WordNet;
- to achieve maximal compatibility across the different resources;

- to build the WordNets relatively independently (re)-using existing resources;

The most important difference of Dravidian WordNet with respect to a language specific WordNet is its multilinguality, which however also raises some fundamental questions with respect to the status of the monolingual information in the WordNets. In principle, multilinguality will be achieved by adding an equivalence relation for each synset in a language to the closest synset in Tamil WordNet. Synsets linked to the same Tamil WordNet synset will be supposed to be equivalent or close in meaning and can then be compared. However, we have to take into consideration the differences across the WordNets. If 'equivalent' words are related in different ways in the different resources, we have to make a decision about the legitimacy of these differences.

In Dravidian WordNet, we will take the position that it must be possible to reflect such differences in lexical semantic relations. The WordNets are seen as linguistic ontologies rather than ontologies for making inferences only. In an inference-based ontology it may be the case that a particular level or structuring is required to achieve a better control or performance, or a more compact and coherent structure. For this purpose it may be necessary to introduce artificial levels for concepts which are not lexicalized in a language or it may be necessary to neglect levels that are lexicalized but not relevant for the purpose of the ontology. A linguistic ontology, on the other hand, exactly reflects the lexicalization and the relations between the words in a language. It is a "WordNet" in the true sense of the word and therefore captures valuable information about conceptualizations that are lexicalized in a language: what is the available stock of words and expressions in a language. In addition to the theoretical motivation there is also a practical motivation for considering the WordNets as autonomous networks. To be more cost-effective, they will be derived (as far as possible) from existing resources, databases and tools. Each site therefore will have different starting points for building their local

WordNet, making it necessary to allow for a maximum of flexibility in producing the WordNets and structures.

3.2 Prototype Implementation

An initial prototype for Tamil WordNet is proposed which can be extended for Dravidian WordNet. In this approach, we try to define and formalize synsets and their relationships using Set theory and first-order logic. A synset is considered as a unique record which consists of the following attributes:

1. Synonyms: words are interchangeable without modifying the context.
2. POS tag: specifies if the synonyms are nouns, verbs, adjectives or adverbs
3. Description
4. Example

The synsets are basically classified into 4 top level sets, namely Nouns, Verbs, Adjectives and Adverbs. Whenever a synset S_i is read into the database by the WordNet interface, the POS tag attribute is checked and the synset is included into the appropriate top level set. Each synset is given a unique id. In parallel, the list of synonyms in the synset is extracted and compiled into a dictionary. While building the dictionary, it is ensured that each word is mapped to list of all the synsets containing that word.

$$D(w) = \{S_1, S_2, \dots, S_n\} \quad n \geq 0$$

We must now establish semantic relations between synsets in each of the top level set. For the Nouns set, we define hyponymy, hypernymy, meronymy, and holonymy. The synsets in the Verb set are related to each other by hyponymy, hypernymy and similarity. The synsets in the Adjectives and Adverbs set relate to each other using similarity. For each top level set, a relationship matrix R can be maintained that captures the semantic relationship between synsets. The rows of R will correspond to the unique synsets in that top level set and the columns of R correspond to the different semantic relations. Each r_{ij} in R can be considered as the set of all synsets S_k that are related to S_i using the semantic relation j .

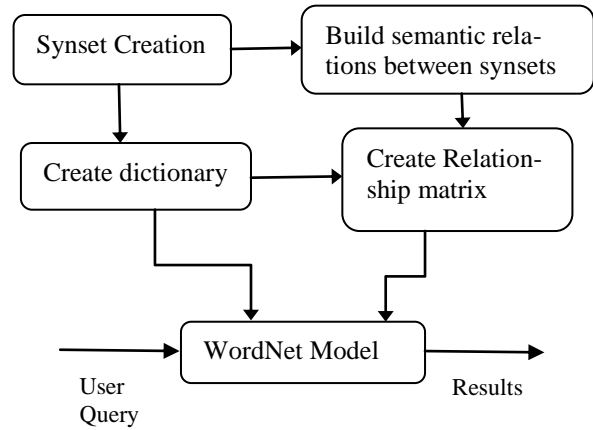
Alternatively, using first order logic, we can define the semantic envelope of a word w as:
 $SE(w_{\text{noun}}) = Hyponym(S_i) \cup Hypernym(S_i) \cup Meronym(S_i) \cup Holonym(S_i)$ where $S_i \in \text{Nouns}$

$$SE(w_{\text{verbs}}) = Hyponym(S_i) \cup Hypernym(S_i) \cup Similar(S_i) \quad S_i \in \text{Verbs}$$

$$SE(w_{\text{adjectives}}) = Similar(S_i) \quad S_i \in \text{Adjectives}$$

$$SE(w_{\text{adverbs}}) = Similar(S_i) \quad S_i \in \text{Adverbs}$$

A simple block diagram of the prototype is shown in Figure 1



Prototype of Wordnet

4 Utilities and Applications

Dravidian WordNet will provide a multilingual network of words linked by semantic relations. This will allow, for example, a search tool to infer other terms, from terms provided by the user, which should be used in searching for information. The network will also provide support across different languages converting terms from one language into other languages. This is particularly useful for users working in a second language and may not have appropriate knowledge of vocabulary. The network will also be used as a basic resource for supporting other applications. The semantic knowledge embodied in the network makes it suitable as a component in expert systems, language translation aids, language learning systems and automatic summarizers.

For application developers the production of a multilingual set of word nets will allow applications to be developed, which can work with a selected language or over a range of languages. This will allow Dravidian developers to compete effectively in the now predominantly (American) English market, providing solutions which work with the user's own language. In addition, it will be possible to create applications which can access and provide information in languages other than the user's own language.

Users will be provided with more effective and powerful tools that they can use in their own language. Applications such as information retrieval will allow non-expert users to access the information they want without the need for familiarity with indexing schemes and terms. Improved access to information will itself provide benefits for European industry, administration and citizens. The support for different languages will reduce language barriers to information whilst retaining the benefits of diversity in expression and culture that the range of Dravidian languages provides.

The resources produced by DWN will have a wide range of users. The current user group, which provides feedback on intermediate project results, comprises members from libraries, software developers, universities and publishers interested in language learning, language generation, machine aided translation, language understanding, information retrieval, electronic publishing and the production of WordNet in additional languages. The end users of the resources will be all those people who utilize the applications that incorporate DWN resources.

5 Conclusion

This paper has attempted to illustrate the need for an independent Dravidian WordNet. As a first step towards building a Dravidian WordNet, a prototype design for Tamil WordNet has also been proposed. WordNet is a natural answer in machine translation systems. It has the potential to interpret source language words and come up with lexical equivalents in the target language in a more natural way as a bilingual does. Building of WordNet is an immediate requirement in the context of information technology equipped with internet in which the web sites in Dravidian languages are getting added up day by day.

References

Alonge, A., N. Calzolari, P. Vossen, L. Bloksma, I. Castellon, T. Marti, W. Peters. 1998. *The Linguistic Design of the EuroWordNet Database*. In: Nancy Ide, Daniel Greenstein, Piek Vossen (eds). Special Issue on EuroWordNet. Computers and the Humanities, Vol 32, Nos. 2-3, 91-115.

- Beckwith, R. and G.A. Miller. 1990. *Implementing a Lexical Network*. International Journal of Lexicography, Vol 3, No.4, 302-312.
- Cruse, D.A. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.
- Fellbaum, C. 1990. *English Verbs as a Semantic Net*. International Journal of Lexicography, Vol 3, No.4, 278-301.
- Fellbaum, C. 1998. *A Semantic Network of English Verbs*. In: Fellbaum, C. (ed.). *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press.
- Fellbaum, C. (ed.) 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- Gross, D. and K.J. Miller 1990. *Adjectives in Wordnet*. International Journal of Lexicography, Vol 3, No.4, 265-277.
- Lyons, J. 1977. *Semantics* (vol.1). Cambridge: Cambridge University Press.
- Miller, G.A.1990. *Nouns in WordNet: a lexical inheritance system*. International Journal of Lexicography Vol 3, No. 4, 245-264.
- Miller, G.A. 1991. *Science of Words*. New York: Scientific American Library.
- Miller, G.A. 1998. *Nouns in WordNet*. In: Fellbaum, C. (ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press
- Miller G.A., R. Beckwith, C. Fellbaum, D. Gross, K.J. Miller. 1990. *Introduction to WordNet: An On-line Lexical Database*. International Journal of Lexicography, Vol 3, No.4, 235-244.
- Miller, K.J. 1998. *Modifiers in WordNet*. In: Fellbaum, C. (ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press
- Nida, E.A. 1975a. *Compositional Analysis of Meaning: An Introduction to Semantic Structure*. The Hague: Mouton
- Nida, E.A. 1975.b. *Exploring Semantic Structure*. The Hague: Mouton
- Pustejovsky, J. 1995. *The Generative Lexicon*. Cambridge: MIT Press.
- Rajendran. S, 1983. *Semantics of Tamil Vocabulary*. (Report of the UGC sponsored Postdoctoral Work in manuscript). Poona: Deccan College Post-Doctoral Research Institute.
- Rajendran. S, 1995. *Towards a Compilation of a Thesaurus for Modern Tamil*. South Asian Language Review 5.1:62-99.
- Rajendran. S, 2001. *taRkaalat tamizc coRkaLanj-ciyam* [Thesaurus for Modern Tamil]. Thanjavur: Tamil University.

- Rajendran, S, 2002. *Preliminaries to the preparation of a Word Net for Tamil*. Language in India 2:1, www.languageinindia.com
- Rajendran, S, 2003. *Pre-requisite for the Preparation of an Electronic Thesaurus for a Text Processor in Indian Languages*. Language in India 3:1, www.languageinindia.com
- Rajendran, S., S. Arulmozi, B. Kumara Shanmugam, S. Baskaran, and S. Thiagarajan. 2002. *Tamil WordNet*. Proceedings of the First International Global WordNet Conference. Mysore: CIIL, 271-274.
- Tengi, R.I. 1998. *Design and Implementation of the WordNet Lexical Database and Searching Software*. In: Fellbaum, C. (ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press.
- Vossen P. (eds.) 1999. *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Dordrecht: Kluwer Academic Publishers,.
- Vossen P. 1999 fc., *EuroWordNet as a multilingual database*. In: Wolfgang Teubert (ed). Berlin: Mouton Gruyter.